Epidemic Marketplace: An Information Management System for Epidemiological Data.

Luis F. Lopes, Fabrício A. B. Silva, Francisco Couto, João Zamite, Hugo Ferreira, Carla Sousa, and Mário J. Silva

University of Lisbon, Faculty of Sciences, LASIGE
epiwork@lasige.di.fc.ul.pt,
WWW home page: http://epiwork.di.fc.ul.pt

Abstract. The Epidemic Marketplace is part of a computational framework for organizing data for epidemic modeling and forecasting. It is a distributed data management platform where epidemiological data can be stored, managed and made available to the scientific community. It includes tools for the automatic interaction with other applications through web services, for the collection of epidemiological data from internet social networks and for discussion of related topics. This paper defines its requirements, architecture and implementation plan based on open-source software. This platform will assist epidemiologists and public health scientists in finding, sharing and exchanging data.

1 Introduction

Internet technologies are becoming essential tools in Epidemiology [1]. Strong computational competencies are necessary for the management and processing of huge quantities of data being produced nowadays by epidemiological studies. Data storage and management requires powerful and stable informatics platforms as well as high speed connectivity for data sharing in local networks and in the Internet. On the other hand, epidemiological modeling requires access to data and extreme processing capacity [2] [3].

The early detection of infectious disease outbreaks is fundamental for the efficient intervention of public health authorities and for the application of disease control measures [4]. So, there is a constant quest for new surveillance methods, capable of decreasing the gap between disease outbreak and its detection [5].

In the last years the internet has been increasingly regarded as a tool for the identification of epidemiological trends and for the fast detection of epidemic outbreaks. There are several systems that provide epidemiological insight based on data gathered in the internet, with sources such as official alerts or news sites [6], search engine queries [7], or even voluntary sentinels [8].

The development of new internet-based epidemiological surveillance systems, has contributed to the continuous increase of existing epidemiological data. However, those data are heterogeneous and organized in a way that makes it hard to share those data among scientists and health professionals.

In this paper, we present the Epidemic Marketplace, a platform for data integration, management and sharing. As a repository for epidemiological data and as a discussion venue, the Epidemic Marketplace will enhance knowledge dissemination and foster data sharing and collaboration among the scientific community. Moreover, this platform is designed to integrate with epidemiological modeling and forecasting computational platforms. This will contribute to the development of a new generation of powerful tools for epidemiological forecast and disease study.

In this paper, we start by reviewing some of the related work on repositories of epidemiological resources and discuss the role of metadata annotation in these systems. We then present the Epidemic Marketplace requirements and architecture. In Section 4, we describe in more detail the Epidemic Marketplace and its modules. In Section 5, we detail the metadata model of the Epidemic Marketplace, and the role of encoding schemes and ontologies for a consistent annotation of metadata. In the following section, we describe the Epidemic Marketplace metadata model design in more detail, with a focus on its specific metadata elements. In Section 7, we discuss strategies to assist metadata content creation. Finally, in Section 8, we present the conclusions and future work.

2 Related Work

There are a large number of health and epidemic data repositories available in the internet. The United States National Library of Medicine presents a list of resources available on the web that provide access to epidemiological data and statistics [9]. One of these services is the CDC Wonder, which provides access to CDC's data in the public domain [10].

MyPubliHealth is a repository of public health resources that uses a metadata schema to organize and manage resources, making information finding easier and faster [11].

A repository of clinical trial data is made available by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), presenting a metadata catalogue describing the stored contents [12]. However, the metadata schema of this repository relies in few fields that are annotated mostly with free text, making metadata contained there too unstructured and semantically poor.

The HEAL (Health Education Assets Library) multimedia repository contains health-related data, such as teaching resources for health sciences education. It is designed to facilitate the sharing of high-quality, freely available multimedia resources located on the HEAL server or in other remote servers [13]. HEAL also functions as a publishing venue, for authors to submit multimedia resources for review and publication there. This repository relies on a detailed metadata schema, defined using selected elements from the IMS Meta-Data Version 1.2.1 specification [14], and several elements defined locally, specific for health sciences. This type of metadata sheme, more structured and relying less is free text, is closer to what we propose for the Epidemic Marketplace.

3 Epidemic Marketplace Architecture

The Epidemic Marketplace is envisioned as a geographic distributed network of interconnected data management nodes, sharing common canonical data models, authorization infrastructure and access interfaces. The distributed architecture should provide improved data access performance, improved availability and fault-tolerance.

Data can be either stored in one or more repositories or retrieved from external data sources using authorization credentials provided by clients. Data can also be replicated among repositories to improve access time, availability and fault tolerance. However, data replication is not mandatory; in several cases data must be stored in a single site due to, for instance, security constraints. It is worth noting, though, that any individual repository that composes the Marketplace will enable virtualized access to these data, once a user provides adequate security credentials.

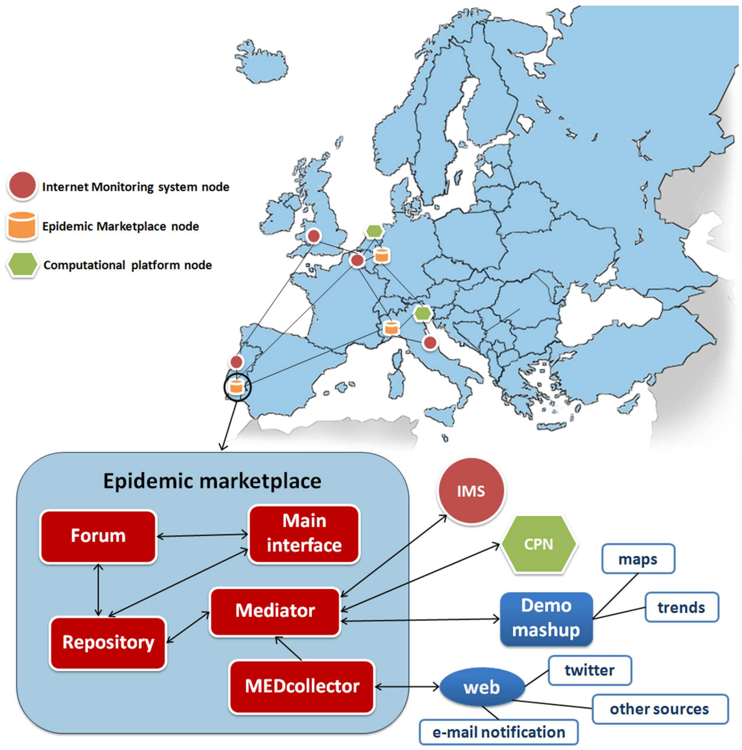

Fig. 1. Example of an envisioned deployment of the Epidemic Marketplace.

According to Fig. 1, each Epidemic Marketplace node will be composed of four modules [15]: 1) a *Digital Repository*, where data will be stored, managed

and made available to the community; 2) the *Mediator*, is responsible for the communication among the different modules and with foreign applications; 3) the *MEDCollector*, which allows the harvesting of data from the web, storing it locally, through the use of workflows; 4) the *Forum*, is a venue for discussion about epidemiological topics, resources stored in the repository, such as datasets, and for the establishment of contacts among the user community.

3.1 Epidemic Marketplace Requirements

Several requirements have been identified when defining the architecture of the Epidemic Marketplace. As a data management platform it needs to:

- **Support the sharing and management of epidemiological data sets:** Registered users should be able to upload annotated datasets, and a dataset rating assessment mechanism should be available. The annotated data set will then compose a catalogue that will be available for users to browse, search and download according to specific permissions.
- **Support the seamless integration of multiple heterogeneous data sources:** Users should be able to have a unified view of related data sources through a common interface. Data should be available from streaming, static and dynamic sources.
- **Distributed Architecture:** The Epidemic Marketplace should implement a geographically distributed architecture deployed in several sites. The distributed architecture should provide improved data access performance, improved availability and fault-tolerance.
- **Support secure access to data:** Access to data should be controlled. The marketplace should provide single sign-on capability, distributed federated authorization and multiple access policies, customizable by users.
- **Modularity:** Not all sites where the Epidemic Marketplace will be deployed need to have all modules installed. Also, some sites may choose to implement different modules to provide new services.
- **Provide data for analysis and simulation in grid environments:** The Epidemic Marketplace will provide data for analysis and simulation services in a grid environment. Therefore, the Epidemic Marketplace should operate seamlessly with grid-specific services, such as grid security services, information services and resource allocation services.
- **Workflow:** The platform should provide workflow support for data processing and external service interaction. This requirement is particularly important for those services that retrieve data from the Epidemic Marketplace, process it, and store the processed data back in the marketplace, such as grid-enabled data analysis and simulation services.
- **Support the creation of a virtual community for epidemic research:** The platform will serve as a forum for discussion that will guide the community into uncovering the necessities of sharing data between providers and modellers. Users will become active participants, generating information and providing data for sharing and collaborating online.

Also, some non-functional requirements were identified:

- **Interoperability:** The Epidemic Marketplace must interoperate with other software. Its design must take into account that, in the future, systems developed by other researchers across the world may need to query the Epidemic Marketplace catalogue and access datasets stored there.
- **Standards-based:** To guarantee software interoperability and the seamless integration of all geographically dispersed sites of the Epidemic Marketplace, the system will be entirely built over web services, authentication and metadata standards.
- **Open-source:** Software packages to be used in the implementation and deployment of the Epidemic Marketplace should be open source, as well as the new modules developed specifically for the Epidemic Marketplace. An open-source based solution reduces development cost, improves software trustworthiness and reliability and simplifies support.

4 Implementation and Deployment

A first prototype of the Epidemic Marketplace has been deployed and is being thoroughly tested since it must be prepared to handle large amount of data and intensive user activity.

Digital Repository The repository was built using Fedora Commons [16], with Muradora as a web frontend [17]. It also relies on other middleware, such as OpenLDAP which is be used for storing user registration an login data.

Fig. 2 depicts a snapshot of the Digital Repository frontend implemented for the Epidemic Marketplace, where users can browse availble collections.

An important architectural feature of the repository is a clear separation between data and metadata [18], since metadata may contain information not directly available in the datasets. While the data organization may be very variable, a well defined metadata schema should be used for data annotation. To complement that, the metadata schema should be based in standards. Being a widely used metadata standard in the Internet, and supported by several standard institutions, Dublin Core (DC) was adopted [19]. Furthermore, metadata must be consistent and semantically meaningful, employing encoding schemes and ontologies [20] [21].

In this work we are developing a metadata schema, which is based upon the DC standard and extended with metadata elements specific for epidemiological resources. This will enable a more meaningful annotation of the datasets contained in the repository. Moreover, ontologies such as the Unified Medical Language System (UMLS) [22], or the Open Biomedical Ontologies (OBO) [23], will be used for the creation of metadata content.

The repository is being populated with annotated datasets and descriptions of epidemiological resources. Some examples are: a dataset of cumulative cases of H1N1; datasets of Twitter messages containin references to diseases, obtained

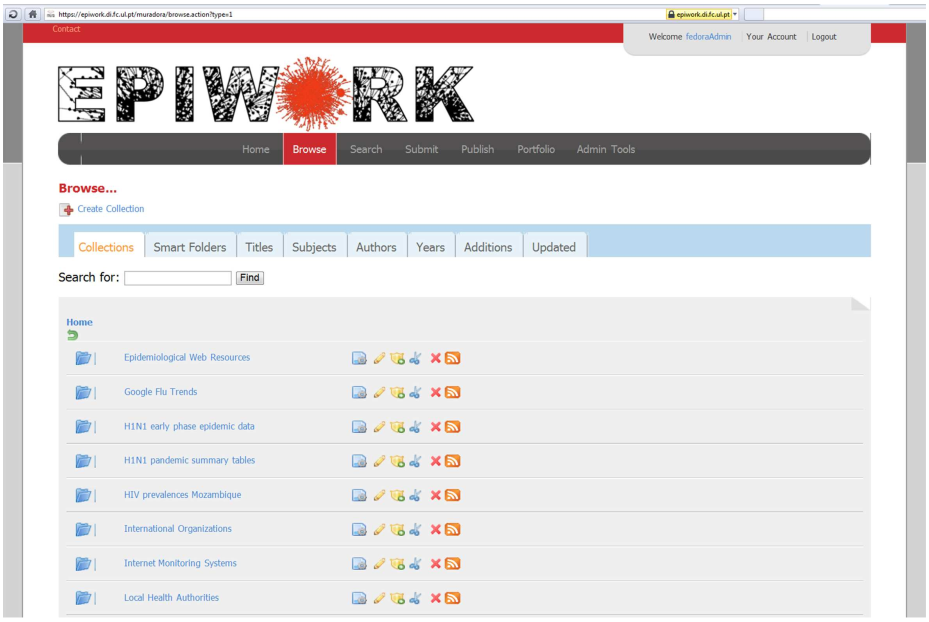

Fig. 2. A section of the repository, where collections may be browsed.

using the Epidemmic Marketplace Data Collector; a dataset with traffic information concerning the main 500 US airports.

Mediator The Mediator is reponsible for the communication with foreign applications and among different Epidemic Marketplace modules. It will communicate with: 1) clients, which retrieve the data collections of the Epidemic Marketplace and produce dynamical trends graphs or maps with goreferenced data; 2) EPIWORK applications, such as Internet-based Monitoring Systems or computational platforms for epidemiological modeling, which can retrieve and store data in the Epidemic marketplace 3) other modules of the Epidemic Marketplace, for example allowing the MEDCollector [24] to automatically store datasets in the Digital Repository.

The Mediator must be able to manage the access to heterogeneous data from different sources, for different diseases, and in different formats, through either query or search interfaces. It is endowed with a RESTful interface, through which clients are able to search and query datasets and corresponding metadata. This interface obeys the principles of REST architecture [25], offering a simple, flexible and intuitive interaction.

The Mediator implements the OAI-PMH protocol [26], which enables the access to the repository metadata by external metadata harvesting applications. It will also follow the OAI-ORE standard guidelines [27], which defines rules for

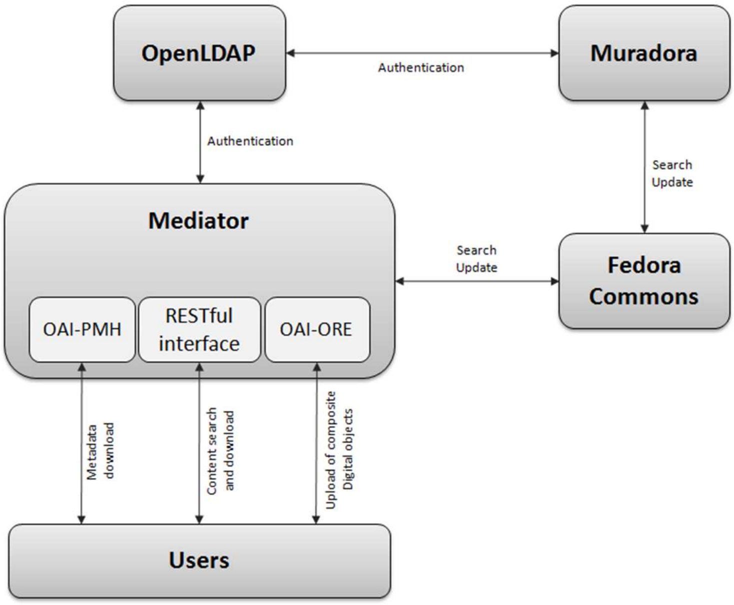

Fig. 3. Diagram of the mediator interactions

the description and interchange of aggregated web resources. These aggregations may combine different types of resources in digital objects.

Fig. 3 represents a diagram of the principal interactions in which the Mediator is involved. According to the diagram, the Mediator will login the users through the information contained in a LDAP directory. The mediator will search and update the information contained in Fedora Commons using methods described in its API. The RESTful services will allow users to search and download of contents.

MEDCollector Data sources such as internet based social networks and search engine query data, may present early evidence of an infection event and propagation [7]. Given the increasing popularity of social networks we can find a large amount of personal information in real time, which may be used for the early detection of epidemic events.

The objective of the MEDCollector is to take advantage of internet data sources and of the large flow of users that currently use social network services, extracting epidemiologic information from there. This module actively collecting messages from Twitter, containing references to specific diseases and locations [28]. New features are being added to this module in order to support the seamless integration of multiple heterogeneous data sources available on the internet, therefore taking not only advantage of social networks but also from a vaster array of epidemiological web data sources. It is also able to passively collect data,

receiving information from other data providers, which can be done through newsfeed or e-mail subscription [24].

The data collected by this module will be stored locally in the Epidemic Marketplace. This is important, since collected data may not be available from its original source after a predefined amount of time. Data are stored in a dedicated database and can be used to create datasets.

Forum The Epidemic Marketplace will serve as an information exchange platform, connecting modelers who search for data, for deriving their models, and those who have data and are searching for the help of modelers to interprete it. This will improve collaborations by direct trustful sharing of data within the communities. The current implementation of the forum is based on phpBB and shares authentication data with the Digital Repository, through LDAP.

5 Metadata Model

There are several metadata standards that can be used to develop metadata schemas, such as ISO/IEC 11179, a standard for storing organizational metadata in a controlled environment [29], and the Dublin Core Metadata Element Set (DC), another metadata standard used for internet resource description [30].

To keep data consistent, is necessary to use not only metadata representation standards for the design of the metadata schema, but also standards for metadata creation, making use of encoding schemes and ontologies.

The metadata model will be essential for the integration of the whole platform, for the interoperability between the different Epidemic Marketplace modules and with other applications. It will be a powerful tool for data manipulation, management and exchange.

5.1 Metadata Model Requirements

We have identified a set of requirements for the creation of a metadata model:

i) **Support well structured and meaningful metadata.** It is important to understand what kind of information is important to the target user and make sure that such information is well described by metadata.

ii) **Straightforward and easy to understand.** This is essential for the system to be accepted by the community, since the creation of metadata may be a complex and time consuming task.

iii) **Use standards to support metadata creation.** Metadata creation should rely the most possible in encoding schemes and ontologies [20] [21], avoiding whenever possible the use of unstructured free text. This makes metadata more consistent, improves information finding and machine readability.

5.2 Encoding Schemes and Ontologies

The use of encoding schemes and ontologies is essential to keep metadata simple and standardized, improving searchability and machine readability.

Encoding schemes make everyone using the same term to refer to a specific attribute. In the Epidemic Marketplace metadata model system several encoding schemes will be used, some of them proposed by Dublin Core, such as the W3C date and time format standard [31], or the RCF4646 standard that defines language codes [32].

Ontologies and databases may be used for the creation of controlled vocabularies to be used in data anotation. Existing ontologies, such as the Unified Medical Language System (UMLS) [22], or other ontologies from the OBO [23], are essential for the annotation of epidemiological and biomedical data. The International Classification of Diseases (ICD) [33], which is integrated in the UMLS [34] will be used to encode disease name. Controlled vocabularies are also useful to standardize other biomedical/epidemiological relevant concepts, providing encoding schemes for drugs [35] or medical procedures [36]. A geographic ontology, with World coverage, is also a primary need [37]. For example, to produce a standard annotation of the spatial coverage the GeoNames database [38] is being looked into as an information source.

6 Metadata Model Design

The metadata model being developed for the Epidemic Marketplace, is based on the DC standard, more specifically in the more recent DCTERMS specification defined by the Dublin Core Metadata Initiative (DCMI) Usage Board [39]. It uses several terms, from the DCTERMS namespace, and implements new ones specific for the annotation of epidemiological data. However, epidemiologic relevant data may cover different areas of knowledge, such as geographic, environmental or demographic, among others.

Table 1 describes specific metadata elements included in the Epidemic Marketplace application profile. Some of these elements may be useful only for specific type of datasets, according to the type of data contained there. For example, if the dataset contains epidemic data then probably metadata elements like Epidemic and Disease should be filled. If it is geographic then the Geographic element should be filled, identifying the type of geographic data in the dataset.

To test this metadata model it is important to access and annotate the most varied possible batch of datasets. This will help to reveal how well the metadata model is fully capable of describing different types of data. Since such batch of datasets is not yet available, we surveyed several epidemiological papers in order to identify types of data used there. We have derived datasets based on the data presentedused in some of those studies and then annotated these pseudodatasets to evaluate how well these concepts mapped to the metadata model. The example, presented in Fig. 4, was based on this approach using a study by Cohen and coworkers [40]. This allowed the identification of important metadata elements

Table 1. Specific metadata terms defined in the Epidemic Marketplace metadata model. The filling of these fields in not mandatory but highly recommendable, if applicable.

Metadata element	Description
Demographic	To describe the type of demographic elements contained in the resource
Disease	To annotate the unequivocal name of a disease that is covered by the resource
Drug	Describe chemical compounds used for disease treatment
Environmental	To describe environmental data in the dataset
Host	Identifies the species of the organism that is the disease host
HostGroup	Identifies a larger taxonomic (or other) group of hosts covered
Geographic	Type of geographic data contained in the resource
Pathogen	Identifies a pathogenic species
PathoGroup	Identifies a taxonomic (or other) group of pathogens
Socio-Economic	Describe social or economic data.
Vaccine	Annotate a vaccine that is covered in the datasete
Vector	Identifies a disease vector organism covered by the dataset

to be included in this model. As the Epidemic Marketplace repository is populated, a wide range of datasets will be available, allowing the testing, evaluation and improvement of this model.

7 Mechanisms to Assist Metadata Creation

Mechanisms to support the user in the process of metadata creation are essential to improve the annotation process, making it faster and easier.

First of all, users must be able to have one or more persistant metadata profiles, which they could modify, save and recover whenever needed. This feature is extremelly useful when it is necessary to annotate more than one similar datasets, allowing the re-use of metadata.

Auto-filling is an essential feature to assist metadata creation. The existing implementation is able to search specific information, inserting it in the metadata form, such as the user name and organization. Other information is also inserted automatically, such as the language, title (based on the file name), date and type. All the automatically inserted metadata can be edited by the user if deemed necessary.

Another important implemented feature are dropdown menus, so the user can select options instead of writing. Dropdown menus will be populated with controlled vocabularies, produced from sources such as thesaurus or ontologies. This will help to keep metadata simple and consistent, using specific encoding schemes.

```
<EM:DISEASE>MALARIA</EM:DISEASE>
<EM:ENVIRONMENT>HUMIDITY</EM:ENVIRONMENT>
<EM:HOST>HUMAN</EM:HOST>
<EM:PATHOGEN>PLASMODIUM FALCIPARUM </EM:PATHOGEN>
<EM:GEOGRAPHIC>TOPOGRAPHIC MAP</EM:GEOGRAPHIC>
<EM:VECTOR>ANOPHELES</EM:VECTOR>
<EM:DEMOGRAPHIC>ECONOMIC ACTIVITY</EM:DEMOGRAPHIC>
<EM:DEMOGRAPHIC>POPULATION SIZE</EM:DEMOGRAPHIC>
<EM:DEMOGRAPHIC>BIRTH RATE</EM:DEMOGRAPHIC>
<EM:DEMOGRAPHIC>DEATH RATE</EM:DEMOGRAPHIC>
<EM:DEMOGRAPHIC>MIGRATION</EM:DEMOGRAPHIC>
```

Fig. 4. Example annotation of data.

8 Conclusion

The prototype of the Epidemic Marketplace has been developed and is now available to the Epiwork project collaborators. The system is being stress tested and some improvements are being implemented. We intend to open the platform to the general public later this year, meanwhile news and announcements can be followed at http://epiwork.di.fc.ul.pt/.

The Digital Repository is fully functional, as well as the forum. Also, there is a prototype of the Mediator, however, only part of the functions to be provided are yet implemented. Furthermore, a prototype of the MEDCollector is also available [24].

A first version of the metadata model to be used in the Epidemic Marketplace has been developed. This metadata model is based on the DC metadata standard, which was designed for simplicity and to describe resources in the internet. The DC schema was initially a vocabulary based on 15 properties. However, a recent revision has made it evolve to follow current best practice for machine-processable metadata [39]. Moreover, the DCMI has provided guidelines for the preparation of Dublin Core Application Profiles (DCAP), which can use elements from the DCTERMS namespace and from other namespaces [41]. This allows the adaptation of DCTERMS based metadata models to very specific applications.

The Epidemic Marketplace metadata model uses elements from DCTERMS for general resource description and in addition uses specific elements for the description of epidemiological data. The inclusion of these epidemiological specific metadata elements, in the metadata schema, will allow a much better structuring of the metadata and an easier implementation of encoding schemes and

ontologies. This metadata model will be common to all Epidemic Marketplace modules, where it will be used to describe, exchange and manipulate data. It will improve interoperability between different modules and with external applications that need to search, access and deposit annotated data in the Epidemic Marketplace.

An issue that needs to be overcome for the success a metadata schema is the time expended in the creation of metadata. A way to avoid this problem is having a simple and lean metadata schema. However, if the schema is too simple it might miss important concepts, so it is necessary to reach an equilibrium between simplicity and the necessity to include key information. Another way to keep the metadata creation process fast and simple is to implement automated aids, to help the user fill in the metadata.

An important factor to consider in the development of a metadata model for the description of epidemiological data is its heterogeneity, since it includes data from many different fields of study, such as behavioral, demographic or geographic data. Furthermore, the variabily of data sources and provenance, study design and collection settings add to this variability. Finally, the metadata model needs to be tested and evaluated, in order to access its efficacy and ease of use. This will allow the review and improvement of the model.

In the future the Epidemic Marketplace aims to provide a framework for the creation and development of epidemiological ontologies, openly addressing the needs of this community and fostering its active involvement. Our goal is to contribute to making ontologies widely accepted by the Epidemiological community and ensuring their sustainable evolution, by replicating the success of similar initiatives, such as the Gene Ontology in Molecular Biology [42].

9 Aknowledgments

The authors want to thank the European Commission for the financial support of the EPIWORK project under the Seventh Framework Programme (Grant #231807), the EPIWORK project partners, CMU-Portugal partnership and FCT (Portuguese research funding agency) for its LaSIGE Multi-annual support.

References

1. S. M. Duke-sylvester, E. N. Perencevich, J. P. Furuno, L. A. Real, and H. Gaff, “Advancing epidemiological science through computational modeling : a review with novel examples,” *Ann. Bot. Fennici*, vol. 45, pp. 385–401, 2008.
2. D. S. Burke, J. M. Epstein, D. a. T. Cummings, J. I. Parker, K. C. Cline, R. M. Singa, and S. Chakravarty, “Individual-based computational modeling of smallpox epidemic control strategies,” *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*, vol. 13, no. 11, pp. 1142–9, 2006.
3. G. V. Bobashev, D. M. Goedecke, F. Yu, and J. M. Epstein, “A hybrid epidemic model: combining the advantages of agent-based and equation-based approaches.”

in *Proceedings of the 2007 Winter Simulation Conference*, S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew and R. R. Barton, Eds., 2007, pp. 1532–1537.
4. K. Wilson and J. S. Brownstein, “Early detection of disease outbreaks using the Internet.” *CMAJ : Canadian Medical Association journal = journal de l’Association medicale canadienne*, vol. 180, no. 8, pp. 829–31, 2009.
5. M. M. Wagner, F. C. Tsui, J. U. Espino, V. M. Dato, D. F. Sittig, R. A. Caruana, L. F. McGinnis, D. W. Deerfield, M. J. Druzdzel, and D. B. Fridsma, “The emerging science of very early detection of disease outbreaks.” *Journal of public health management and practice : JPHMP*, vol. 7, no. 6, pp. 51–9, 2001.
6. C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein, “HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports.” *Journal of the American Medical Informatics Association : JAMIA*, vol. 15, no. 2, pp. 150–7.
7. J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, “Detecting influenza epidemics using search engine query data.” *Nature*, vol. 457, no. 7232, pp. 1012–4, 2009.
8. S. P. van Noort, M. Muehlen, H. Rebelo de Andrade, C. Koppeschaar, J. M. Lima Lourenço, and M. G. M. Gomes, “Gripenet: an internet-based system to monitor influenza-like illness uniformly across Europe.” *Euro surveillance : bulletin européen sur les maladies transmissibles = European communicable disease bulletin*, vol. 12, no. 7, pp. E5–6, 2007.
9. National Information Center on Health Services Research and Health Care Technology (NICHSR) - HSR Information Central. [Online]. Available: http://www.nlm.nih.gov/hsrinfo/datasites.html
10. CDC Wonder. [Online]. Available: http://wonder.cdc.gov/
11. D. Revere, P. Bugni, and S. Fuller, “A Public Health Knowledge Management Repository that Includes Grey Literature,” *Publishing Research Quarterly*, no. 1, pp. 65–70, 2007.
12. A. J. Cuticchia, P. Cooley, R. D. Hall, and Y. Qin, “NIDDK data repository: a central collection of clinical trial data.” *BMC medical informatics and decision making*, vol. 6, no. 1, p. 19, 2006.
13. Heal. (2005) Heal Metadata Elements Description, version 1.6. [Online] Available: http://www.healcentral.org/services/schema/HEALmdElementsDescript_v1p6.pdf.
14. IMS Global Learning Consortium. (2005) IMS Application Profile Guidelines overview. [Online]. Available: http://www.imsglobal.org/ap/apv1p0/imsap_oviewv1p0.html
15. M. J. Silva, F. A. B. da Silva, L. F. Lopes, and F. M. Couto, “Building a Digital Library for Epidemic Modelling,” in *Proceedings of ICDL 2010 - The International Conference on Digital Libraries*, vol. 1. New Delhi, India: TERI Press – New Delhi, India, 2010.
16. C. Lagoze, S. Payette, E. Shin, and C. Wilper, “Fedora: an architecture for complex objects and their relationships,” *International Journal on Digital Libraries*, vol. 6, no. 2, 2006.
17. C. Nguyen and J. Dalziel, “Muradora: A Turnkey Fedora GUI Supporting Heterogeneous Metadata, Federated Identity, And Flexible Access Control,” in *Third International Conference on Open Repositories 2008*, 2008.
18. E. Stolte, C. von Praun, G. Alonso, and T. Gross, “Scientific data repositories:designing for a moving target,” *International Conference on Management of Data*, 2003.

19. DCMI. The Dublin Core Metadata Initiative. [Online] Available: http://dublincore.org/. [Accessed April, 2010].
20. A. Goni, E. Mena, and A. Illarramendi, "Querying Heterogeneous and Distributed Data Repositories using Ontologies," in *Proceedings of the 7th European-Japanese Conference on Information Modelling and Knowledge Bases*, 1997.
21. P. Fox, D. McGuinness, D. Middleton, L. Cinquini, J. Darnell, J. Garcia, P. West, J. Benedict, and S. Solomon, *The Semantic Web - ISWC 2006*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, vol. 4273.
22. O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology." *Nucleic acids research*, vol. 32, no. Database issue, pp. D267–70, 2004.
23. B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration." *Nature biotechnology*, vol. 25, no. 11, pp. 1251–5, 2007.
24. J. Zamite, F. Silva, F. Couto, and M. J. Silva, "MEDCollector: Multisource epidemic data collector," in *Proceedings of the 1st International Conference on Information Technology in Bio- and Medical Informatics. DEXA 2010*, 2010.
25. R. T. Fielding, "Architectural styles and the design of network-based software architectures," PhD Thesis, University of California, Irvine, 2000.
26. OAI. (2008) The Open Archives Initiative Protocol for Metadata Harvesting. [Online]. Available: http://www.openarchives.org/OAI/openarchivesprotocol.html
27. D. Tarrant, B. O'Steen, T. Brody, S. Hitchcock, N. Jefferies, and L. Carr, "Using OAI-ORE to Transform Digital Repositories into Interoperable Storage and Services Applications," *The Code4Lib Journal*, vol. 6, 2009.
28. L. F. Lopes, J. M. Zamite, B. C. Tavares, F. M. Couto, F. Silva, and M. J. Silva, "Automated Social Network Epidemic Data Collector." in *INForum informatics symposium*, Lisbon, 2009.
29. ISO/IEC, "International Standard: Information technology - Metadata," Switzerland, 2004.
30. D. Powell, M. Nilsson, A. Naeve, P. Johnston, and T. Baker. (2007) DCMI Abstract Model. [Online]. Available: http://dublincore.org/documents/abstract-model/
31. M. Wolf and C. Wicksteed. (1998) W3C Date and Time Formats. [Online]. Available: http://www.w3.org/TR/NOTE-datetime
32. A. Phillips and M. Davis. (2006) Best Current Practice - Tags for Identifying Languages. [Online]. Available: http://www.ietf.org/rfc/rfc4646.txt
33. WHO. (2007) ICD-10. International Statistical Classification of Diseases and Related Health Problems. 10th Revision. [Online]. Available: http://apps.who.int/classifications/apps/icd/icd10online/
34. M. Schopen and S. J. Nelson, "ICD-10 and the Unified Medical Language System (UMLS) Recommendations," Brisbane, Australia, 2002. [Online]. Available: http://www.dimdi.de/static/en/klassi/koop/who/etc/etc_02_44.pdf
35. G. H. Merrill, P. B. Ryan, and J. L. Painter, "Construction and Annotation of a UMLS / SNOMED-based Drug Ontology for Observational Pharmacovigilance," *Methods*, 2008.
36. WHO. International Classification of Health Interventions (ICHI). [Accessed April, 2010]. [Online]. Available: http://www.who.int/classifications/ichi/en/

37. M. S. Chaves, M. J. Silva, and B. Martins. (2005) A Geographic Knowledge Base for Semantic Web Applications. [Online] Available: http://www.lbd.dcc.ufmg.br:8080/colecoes/sbbd/2005/003.pdf.
38. GeoNames Team. GeoNames. [Online] Available: http://www.geonames.org/. [Accessed April, 2010].
39. DCMI Usage Board. (2008) DCMI Metadata Terms. [Online] Available: http://dublincore.org/documents/2008/01/14/dcmi-terms/.
40. J. M. Cohen, K. C. Ernst, K. A. Lindblade, J. M. Vulule, C. C. John, and M. L. Wilson, "Topography-derived wetness indices are associated with household-level malaria risk in two communities in the western Kenyan highlands." *Malaria journal*, vol. 7, p. 40, 2008.
41. K. Coyle and T. Baker. (2009) Guidelines for Dublin Core Application Profiles. [Online] Available: http://dublincore.org/documents/2009/05/18/profile-guidelines/.
42. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." *Nature genetics*, vol. 25, no. 1, pp. 25–9, 2000.